UNLIMITED

(2)

RSRE
MEMORANDUM No. 4512

ROYAL SIGNALS & RADAR ESTABLISHMENT

THE USE OF LINEAR DISCRIMINANT ANALYSIS IN THE *ARM* CONTINUOUS SPEECH RECOGNITION SYSTEM

Authors: S M Peeling & K M Ponting

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS.

RSRE MEMORANDUM No. 4512

UNLIMITED

0120200

CONDITIONS OF RELEASE

308900

****

DRIC U

****

DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

Royal Signals and Radar Establishment

Memorandum 4512

The Use of Linear Discriminant Analysis in the *ARM* Continuous Speech Recognition System

S M Peeling and K M Ponting

9th January 1992

Abstract

Linear discriminant analysis is used to generate speech data transformations. Thi- -ansformed data is then used within the *ARM* continuous speech recognitio- -ystem. Experiments are described using transformed data in conjunction with variable frame rate analysis and word transition penalties. Speaker independent results are reported which are as good as the best obtained previously using cosine transformations and variable frame rate analysis. The two sets of results are compared and commented on.

INTENTIONALLY BLANK

Contents

1 Introduction 1

2 Linear Discriminant Analysis 1

3 Experimental Setup 3

4 Results 6

4.1 VFR900 Data 6

4.1.1 Single Frame Transform 6

4.1.2 Three Frame Transform 8

4.1.3 Three Frame Transform With A Second Application Of VFR Analysis 8

4.1.4 Three Frame Transforms With Forced Differences 11

4.2 Full Frame Rate Data 13

4.2.1 Three Frame Transforms 13

4.2.2 Three Frame Transforms With Forced Differences 16

5 Conclusions 18

Appendix A The Use Of Word Transition Penalties 21

Appendix B Comparison Of LDA And MFCC Results 23

List of Figures

1 Principal components of artificial data 2

2 Word errors for three frame transforms on VFR900 data 9

3 The effect of applying VFR analysis for the second time 9

4 Word errors for three frame transforms (with forced differences) on VFR900 data 12

5 Word errors for three frame transforms 15

6 Word errors for three frame transforms with forced differences 17

7 The use of word transition penalties on VFR900 data 22

8 The use of word transition penalties on full rate data 22

List of Tables

1 Recognition results for single frame transforms 7

2 Recognition results for three frame transforms and VFR900 data . . . 8

3 Recognition results from applying VFR analysis for the second time . 10

4 Recognition results for three frame transforms and VFR900 data . . . 11

5 Recognition results for three frame transforms 14

6 Recognition results for three frame transforms with forced differences 16

7 Recognition results for all versions of the SI-ARM system 24

1 Introduction

The work described in this report was conducted at the UK Speech Research Unit. It is partly supported by IED project 3/1/1057 on Speech Recognition Techniques and also forms part of the Airborne Reconnaissance Mission (*ARM*) continuous speech recognition project. The aim of the *ARM* project is accurate recognition of continuously spoken airborne reconnaissance reports using a speech recognition system based on phoneme-level hidden Markov models (HMM). The *ARM* project is described in detail in [14], [15].

The *ARM* system currently applies a discrete cosine transformation to a spectral representation of the speech to produce (so called) mel frequency cepstral coefficients (MFCCs). This linear transformation and representation is commonly used in current speech recognition systems (eg [6], [8]).

Linear discriminant analysis (LDA) can be used to transform data in order to improve a classification system and has the advantage of determining the relative importance of the transformed coefficients in the discrimination process. This allows for some degree of (informed) data reduction. A short description of LDA can be found in Section 2; a more complete explanation can be found in [4], [5].

This paper is a companion to [10] which dealt with the use of LDA for speaker dependent data. Here, the LDA transformation has been applied to speaker independent data in the *ARM* system. Previous papers (eg [12]) have shown that the performance of the *ARM* system can be improved by using VFR analysis and word transition penalties to reduce the numbers of insertions. Results are presented here using the LDA transformation in conjunction with both these techniques.

The results are presented in two sections; in Section 4.1 the LDA transform is applied to data which has previously undergone VFR analysis whilst in Section 4.2 the VFR analysis is applied after the LDA transformation. Most of the results presented used a fixed word transition penalty of 20. Details of the experiments performed to determine this value can be found in Appendix A.

Appendix B attempts to compare the results described here with those obtained using the current *ARM* system described in [14].

2 Linear Discriminant Analysis

This section will give a broad overview of LDA; for a more detailed description see [4], [5].

In any pattern classification task the main objective is to assign some unknown pattern to a particular class. In order to achieve this, it is necessary to

attempt to match one set of features against another. Ideally this set of features should not be too large and there should be some information as to the relative importance of individual features in the classification process.

In speech recognition the cosine transformation is commonly used to improve the discrimination process (and to reduce the number of features in some systems). One motivation for the use of this transformation was given by Pols ([11]). He showed that the first three cosine components were a reasonable approximation to the first three principal components of his speech data.

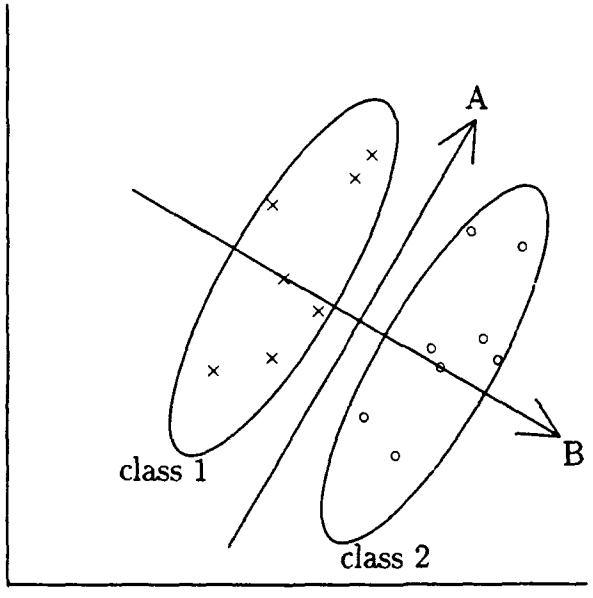

Figure 1: The first two principal components for the two classes of (artificial) data.

However principal components analysis is primarily concerned with the total covariance matrix of the input data and takes no account of any known class labels. Therefore the improvement in discrimination is a by-product of this analysis, rather than its chief aim. This can be seen from the artificial data shown in Figure 1. Principal components analysis will give direction A as the first principal component, and B as the second, but all discrimination relies on B.

Linear discriminant analysis provides a method of examining class-labelled data and determining linear combinations of features which provide maximum discrimination. LDA has the added advantage that these features are ordered so that their relative importance in this discrimination process is known. Because of this, LDA can be used to provide a reliable means of data reduction.

It is worth noting that LDA applied to the data in Figure 1 would give direction B as the first linear discriminant.

Geometrically, the LDA transformation corresponds to a rotation followed by a scaling followed by a rotation of n dimensional space. These are constructed so that variations between the classes are concentrated in the lower dimensions of the space. Hence, the final rotation can be, and usually is, followed by a truncation to reduce dimensionality.

The LDA analysis assumes that the within class covariance matrix is the same for each class and relies only on pooled within (W) and between (B) class covariance matrices. After the transformation, the corresponding W' is the identity matrix and B' is diagonal with the variances down the diagonal ordered by size. This means that the set of features which give the greatest between class discrimination can easily be extracted (ie the less important features can be discarded).

3 Experimental Setup

In all the experiments reported here, the data created was passed to the *ARM* system which is described in [13], [14] and [15]. The version of the *ARM* system used here was a triphone based HMM system with a vocabulary size of approximately 500 words and 1500 triphones.

The speech data used were obtained by passing digitised speech signals through a 27 channel filter bank analyser at 100 frames per second. The filters were spaced on a non-linear frequency scale based on that in [3]. As with the experiments reported in [9], the bottom (DC-60Hz) channel was omitted. Hence only the top 26 channels output from the filter bank were used.

The class labels used for LDA were based on forced alignment of the training data to previously generated HMMs. Each speech frame was given a class label indicating the phoneme and model state within the aligned triphone models. Hence, since most of the models contained three states, there were three classes for each phoneme. These class labels were then used to calculate pooled within class and total covariance matrices; the between class matrix being obtained by subtraction.

Many different transformations can be obtained by using different representations of the filter bank speech data. The simplest is to consider a single frame of data and its associated class. However it can be useful to include information from surrounding frames. For example, three input frames at a time could be considered, with the relevant class determined by the centre frame. In more complicated schemes differences can be incorporated whereby instead of considering surrounding frames directly, the differences between them are used. Similarly, regression coefficients over several frames can be used. In [5], Hunt and Lefébvre report using log filter outputs,

regression coefficients and a notch filter representation as the primary representation to which LDA is applied.

In the experiments reported here, three different schemes for calculating the LDA transformation have been employed. In the first, and simplest case, the analysis considered a single frame at a time and will be referred to as a "single frame transform". In the second case, three input frames were used, with the classification dependent upon the central frame and this is referred to as a "three frame transform". The third case was similar to the second in the number of frames considered and the frame used to determine the classification. However the transform matrix was created from a vector containing the centre frame and the difference between the two outer frames. For example, if the three input frames were $a\,b\,c$ then the vector used to construct the transform matrix would contain b in the first half and $a - c$ in the second half. This is referred to as a "three frame transform with forced differences".

Since one of the properties of the LDA transform is the ordering of the elements in the output vector it was to be expected that some could be discarded. All the experiments reported here discarded some of the elements in the output vectors.

It was shown in [12] that the recognition performance of the *ARM* system could be significantly improved by the use of word transition penalties which were used to control the relative numbers of insertions and deletions. However the results for speaker dependent LDA data reported in [10] did not show the same improvement. Experiments were therefore conducted to investigate the effect of word transition penalties on speaker independent LDA data and these results can be found in Appendix A.

Variable frame rate (VFR) analysis was also used both before and after the LDA transformation. A full description of VFR analysis can be found in [9]. It is sufficient here to state that VFR analysis is a data dependent method of data reduction. The degree of data reduction is determined by the threshold and duplication limit used. Each output frame contains an extra value which is the count of the number of frames represented by that frame. In the VFR experiments, various thresholds were used whilst the duplication limit remained at 50.

Speaker independent recognition experiments were conducted using speech from the "SI89" 321 speaker corpus described in [14]. The data used here consisted of training material from 60 (male) speakers each of whom had recorded three complete *ARM* reports. The test set consisted of three reports from each of ten (different) male speakers. This resulted in a test set containing 1573 words and 6594 phonemes.

Recognition was performed using a one-pass dynamic programming algorithm with beam search and partial traceback [1]. Results are presented in terms of *% words (or phonemes) wrong* and *% word (or phoneme) errors*. These are computed as follows, using dynamic programming to align the true transcription of the test

data with the output of the recogniser:

$$\%\ words\ wrong = \frac{S+D}{N} \times 100,$$
$$\%\ word\ errors = \frac{S+D+I}{N} \times 100$$

where N is the number of words in the test set, and S, D and I are the number of words recognised as incorrect, deleted and inserted respectively[1]. For all the results reported here $N = 1573$.

Recognition results are reported for two levels of syntactic constraint. All the phoneme results come from employing the *phoneme* syntax in which any sequence of triphones can be recognised and the results are scored according to whether or not the correct phoneme is recognised. The word results are obtained from the *word* syntax which allows recognition of any sequence of non-speech sounds and words from the *ARM* vocabulary.

Significance levels for the results presented here are obtained using the matched pairs test suggested in [2] and implemented as described in [7].

Two different sets of results are presented. In the first, the LDA transforms were created using data which had previously undergone VFR analysis. A VFR threshold of 900 had been applied to filter bank data which had reduced the data rate by about 60%. Each data frame consisted of 27 elements [2]. Hence the LDA transforms were square matrices of sizes 27, 81 and 54 for the single frame, three frames and three frames with forced differences respectively. It is worth noting that the model file produced from the cepstral representation of this data was used to produce the class labels for the LDA . This model file gave nearly 73% word accuracy when used with word transition penalties and was the best available at that time ([14]).

In the second set of results, the same model was used to create the class labels which were then made to refer to full frame rate data. In other words, the VFR analysis was "reversed". The data frame in this case contained 26 elements resulting in square transform matrices of sizes 26, 78 and 52 elements for the single frame, three frames and three frames with forced differences respectively. In order to reduce the data rate it was necessary to apply VFR analysis to the transformed data (after elements had been discarded). The VFR count was then appended to the transformed frame.

[1] Previous papers have quoted percentage word accuracy results which are defined as:– $100 - \%word\ errors$.

[2] The DC channel was omitted but the VFR count was appended to each frame.

4 Results

It was reported in [12] that word transition penalties could be used to significantly improve the performance of the *ARM* system. However this was not the case for the speaker dependent LDA experiments reported in [10]. Therefore, initial experiments here investigated the effect of word penalties for the various transform matrices.

The full results from these experiments are shown in Appendix A. From these it was clear that it was possible to improve the word recognition performance by using a word transition penalty of 20. Hence most of the results quoted here use this penalty. However, for comparison purposes, some of the figures also show word errors obtained without word transition penalties.

4.1 VFR900 Data

The LDA transform matrices used here were created from data which had undergone VFR analysis. The analysis had been applied to filter bank data using a threshold of 900 and a duplication limit of 50. This threshold had reduced the data rate by about 60%.

4.1.1 Single Frame Transform

A few experiments were conducted using transform matrices created using a single frame of input data. The transformed data frame contained 27 elements (26 channels from the filter bank plus the VFR count) and initial experiments were conducted to investigate how many of these elements were important in the discrimination process. The results for various numbers of retained/discarded elements are shown in Table 1[3].

The results for the various numbers of retained elements are very similar, and uniformly disappointing. It was shown in [14] that the performance was significantly improved by including information from surrounding frames so it was therefore decided to conduct no further experiments using the single frame transform.

[3]In order to simplify the comparison between the various sets of results, all the tables also show the size of the output vector used in the reestimation process.

No of Elements		% Phone		% Word	
Output Vector	Retained/ Discarded	Wrong	Errors	Wrong	Errors
12	12/15	73.8	73.9	31.7	46.3
17	17/10	71.2	71.3	30.7	46.5
22	22/5	72.3	72.5	30.6	46.6

Table 1: Full recognition results obtained using single frame transform matrices with various numbers of elements retained/discarded and word transition penalties of 20.

4.1.2 Three Frame Transform

For the results quoted in this section, the LDA transform matrix was created by considering three frames of input data. This resulted in an output vector containing 81 elements.

No of Elements		% Phone		% Word	
Output Vector	Retained/ Discarded	Wrong	Errors	Wrong	Errors
11	11/70	66.3	66.5	22.4	31.6
13	13/68	63.2	63.4	21.6	30.8
16	16/65	59.9	60.2	20.7	31.0
19	19/62	57.4	57.7	20.6	30.8
26	26/55	54.5	55.7	21.4	33.8

Table 2: Full recognition results obtained using three frame transform matrices on VFR900 data with various numbers of elements retained/discarded and with word transition penalties of 20.

It was obviously impractical to use the full output vector in the reestimation stage so initial experiments investigated the effect of retaining/discarding differing numbers of elements. The full results are shown in Table 2 and the word errors are summarised in Figure 2.

These results were a great improvement over those obtained using the single frame transform but were no better than the best reported in [14] using MFCC data.

4.1.3 Three Frame Transform With A Second Application Of VFR Analysis

Although the transform matrices had been created using data to which VFR analysis had been applied, it was decided to investigate the effect of a further application of VFR. Since the results reported in [14] used 8 MFCCs plus differences, it was decided to use VFR analysis on the data set with 16 elements retained[4]. The first task was to investigate the effect of different VFR thresholds on the data set and the results (for a typical testing file) are shown in Figure 3.

Previous experience has suggested that a good VFR threshold halves the data rate. However it must be remembered that this data has already undergone VFR analysis and hence the data rate has been reduced by 60% (ie there are 40% of the

[4]This choice arose from a misunderstanding over how many elements were in the output vector produced by 8 MFCCs plus differences.

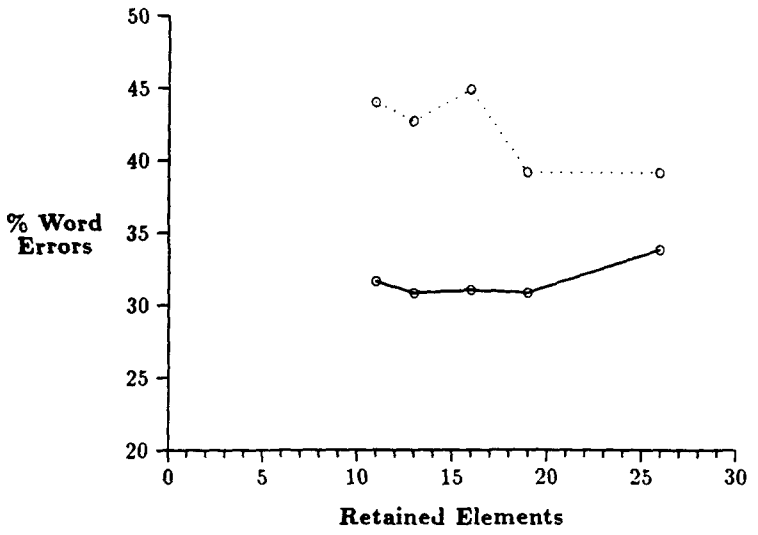

Figure 2: Word errors for various numbers of retained elements, with VFR900 data transformed using three frame transform matrices. The dotted line used no word transition penalties whilst the solid line used penalties of 20.

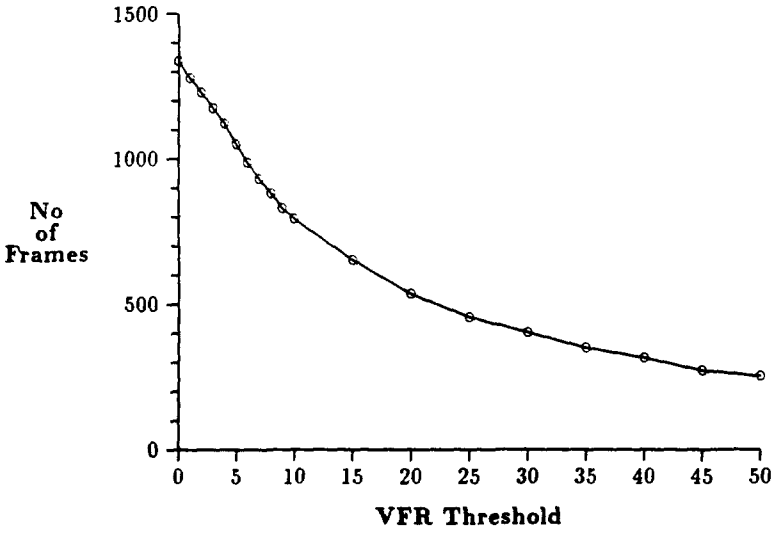

Figure 3: Number of frames processed during testing on three frame transform data (created using VFR900 data) with 16 elements retained, for various VFR thresholds.

original frames remaining). It was therefore decided to try two low thresholds which would result in a slight reduction in data rate. The recognition results are shown in Table 3[5].

Output Vector Size	VFR Threshold	% Frames Remaining	% Phone		% Word	
			Wrong	Errors	Wrong	Errors
16	–	39.1	59.9	60.2	20.7	31.0
17	2	35.9	60.1	60.4	19.8	28.5
17	4	32.8	62.7	62.9	21.0	29.6

Table 3: Full recognition results obtained using three frame transform matrices on VFR900 data to which VFR analysis was applied for the second time with the thresholds shown. Word transition penalties of 20 were used.

It can be seen that the repeated use of VFR analysis has resulted in a marked improvement in performance.

[5]The size of the output vector increases because of the addition of the VFR count.

4.1.4 Three Frame Transforms With Forced Differences

Here, three input frames were considered but the LDA matrix was based on the centre frame and the difference of the surrounding frames. Hence the output vector contained 54 elements.

Various discard factors were used and the results are shown in Table 4.

No of Elements		% Phone		% Word	
Output Vector	Retained/ Discarded	Wrong	Errors	Wrong	Errors
12	12/42	64.3	64.6	20.8	29.7
13	13/41	63.3	63.6	20.1	28.9
15	15/39	60.8	61.4	19.5	28.6
18	18/36	58.4	58.8	20.3	30.2

Table 4: Full recognition results obtained using three frame transform matrices with forced differences on VFR900 data with various numbers of elements retained/discarded and with word transition penalties of 20.

The word errors are summarised in Figure 4. For comparison purposes, the results using three frame transforms are also shown. From this it can be seen that there is a slight performance improvement when using forced differences.

These results are still slightly worse than the best reported in [14].

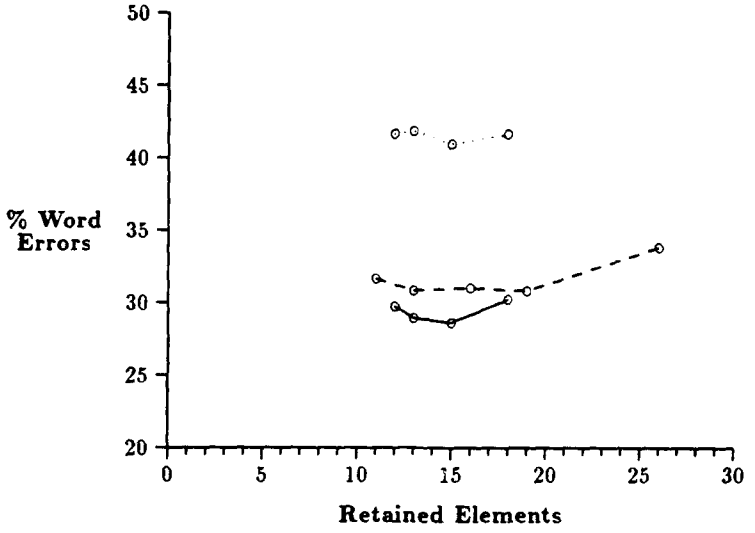

Figure 4: Word errors for various numbers of retained elements, with VFR900 data transformed using three frame transform matrices with forced differences. The dotted line used no word transition penalties whilst the solid line used penalties of 20. The dashed line shows the errors from three frame transform matrices and word transition penalties of 20.

4.2 Full Frame Rate Data

The early results were uniformly disappointing. By using data which had undergone VFR analysis to create the LDA transforms, information had been thrown away. However, it is desirable to base the LDA transform on as much data as possible and thus it would seem better to create the LDA transform matrices using full frame rate data, then apply VFR analysis to this transformed data. The results quoted in this section all come from this system.

In order to reduce the number of parameters to be investigated, the word transition penalty for all these experiments remained set at 20 [6]. It was then possible to concentrate on investigating the effect of discarding different numbers of elements prior to applying various VFR thresholds. All experiments used three frame transforms, with and without forced differences, and the results are presented in the following sections.

It is difficult to compare the different numbers of retained elements and their associated VFR thresholds directly. Hence the results will quote the number of frames remaining in a typical testing file after VFR analysis has been applied. It is then possible to compare results for similar levels of data reduction.

Note that in all the tables in this section the output vector had the VFR count appended and hence its size was one greater than the number of elements retained.

4.2.1 Three Frame Transforms

The transform matrices were created by considering three frames of input data resulting in an output vector containing 78 elements.

As before, experiments were conducted into the effect of retaining varying numbers of elements in the output vectors. In order to reduce the data to a manageable size, it was necessary to employ VFR analysis after discarding elements. Various thresholds were used in an effort to investigate the effect of VFR analysis.

The full recognition results are shown in Table 5 and the word errors summarised in Figure 5.

The overall level of word errors is virtually identical to that shown in Figure 2. The main difference lies in the greater degree of data reduction for the results in Figure 2 where only 40% of the original frames remain.

[6] The results reported in Appendix A show that this is still a suitable value.

No of Elements		VFR	% Frames	% Phone		% Word	
Output Vector	Retained/ Discarded	Threshold	Remaining	Wrong	Errors	Wrong	Errors
7	6/72	1	40.8	78.5	78.5	34.0	46.7
9	8/70	1	51.3	67.6	67.8	23.8	32.7
		2	37.4	72.4	72.8	26.9	37.3
11	10/68	1	61.1	61.5	62.5	20.5	32.5
		2	45.1	65.0	65.5	21.2	29.4
		3	37.5	67.7	68.1	23.0	32.2
13	12/66	2	54.5	59.8	60.7	19.9	28.9
		3	45.0	62.7	63.6	21.2	30.3
		4	39.0	64.7	65.2	21.2	29.9
16	15/63	2	64.2	54.7	56.2	19.7	32.3
		4	48.1	58.4	59.8	19.1	29.6
		6	38.1	61.2	62.0	20.8	30.0
19	18/60	3	61.6	53.5	54.9	19.0	30.8
		5	49.6	55.4	57.0	19.1	29.1
		8	37.9	59.2	60.3	21.2	32.2

Table 5: Full recognition results obtained using three frame transform matrices on full frame rate data data to which VFR analysis was applied with the thresholds shown. Word transition penalties of 20 were used.

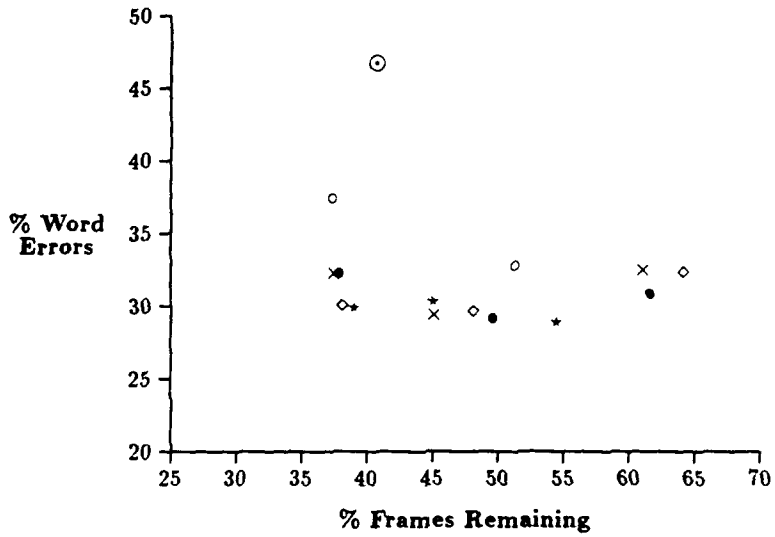

Figure 5: Word errors obtained using three frame transform matrices on full frame rate data data to which VFR analysis was applied to reduce the number of processed frames as shown. Various numbers of retained elements:- (a) 6 - ⊙, (b) 8 - ○, (c) 10 - ×, (d) 12 - ⋆, (e) 15 - ◇ and (f) 18 - ●. Word transition penalties of 20 used.

4.2.2 Three Frame Transforms With Forced Differences

Here, three input frames were considered and forced differences used, whereby the LDA matrix was based on the centre frame and the difference of the surrounding frames. Hence the output vector contained 52 elements.

The full recognition results obtained from retaining various numbers of elements and using VFR analysis, with different thresholds, are shown in Table 6. The word errors are summarised in Figure 6.

No of Elements		VFR	% Frames	% Phone		% Word	
Output Vector	Retained/ Discarded	Threshold	Remaining	Wrong	Errors	Wrong	Errors
13	12/40	2	61.1	58.5	59.5	19.5	29.2
		3	50.1	60.1	60.8	19.9	27.6
		4	38.0	63.0	63.6	20.7	29.0
		5	39.0	64.7	65.1	21.8	30.9
16	15/47	3	60.2	54.9	55.8	18.9	28.4
		4	52.7	57.4	58.8	19.4	31.0
		5	47.2	58.4	59.4	18.6	27.0
		7	39.4	60.8	61.4	20.3	29.6
19	18/34	4	60.1	53.6	54.7	18.1	27.7
		5	53.7	54.6	55.4	17.4	26.1
		7	45.3	56.3	57.5	19.0	28.3
		9	38.9	58.1	58.8	20.4	30.3

Table 6: Full recognition results obtained using three frame transform matrices with forced differences on full frame rate data data to which VFR analysis was applied with the thresholds shown. Word transition penalties of 20 were used.

The best result is obtained by retaining 18 elements in the output vector and reducing the data rate by about 50% (i.e. by using a VFR threshold of 5). This gives 26% word errors which is the best result reported here.

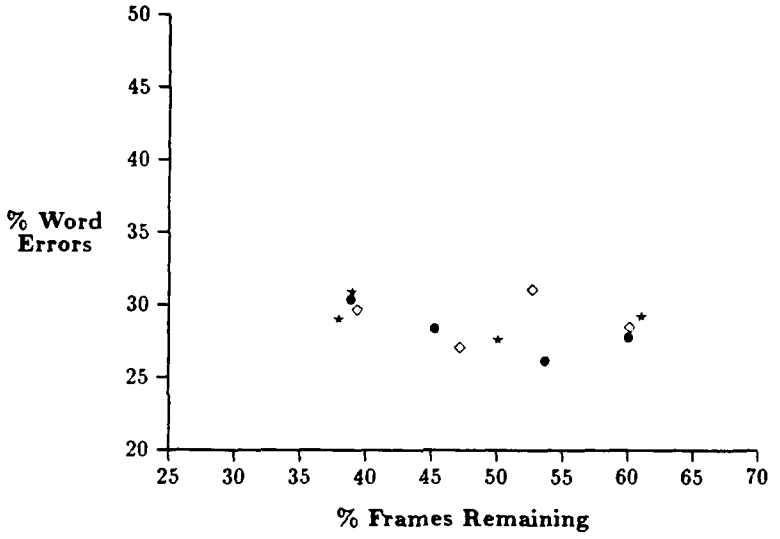

Figure 6: Word errors obtained using three frame transform matrices with forced differences on full frame rate data data to which VFR analysis was applied to reduce the number of processed frames as shown. Various numbers of retained elements:- (a) 12 - $\star$, (b) 15 - $\diamond$ and (c) 18 - $\bullet$. Word transition penalties of 20 used.

5 Conclusions

It must be stressed that this series of experiments was not exhaustive so all conclusions must be viewed as tentative.

It is difficult to do a direct comparison between the results obtained using data which had undergone VFR analysis prior to the calculation of the LDA matrices and those where the matrices were calculated from full frame rate data. In the former case, over half the frames had been removed so when using forced differences the differences for a frame at time t were taken, on average, over frames at $t \pm 2$. However for the full frame rate data these frames were at $t \pm 1$.

A discussion of these results and the best MFCC results can be found in Appendix B. Some comparison of the two sets of results is given.

The (tentative) conclusions for the use of linear discriminant analysis on this speaker independent database are:-

- The results using single frame transforms were much worse than those obtained using three frame transforms, with or without forced differences.
- When using three frame transforms better performance is obtained when forced differences are employed.
- The best results obtained here are on a par with the best reported in [14], i.e about 74% word accuracy.
- It has been shown that when using the full frame rate data to create the LDA transforms, a suitable VFR threshold can be found by reducing the data rate by about 50%.
- It was possible to improve the word accuracy for the three frame transform created using data which had undergone VFR analysis, by a further application of VFR analysis. This probably indicates that the initial choice of VFR threshold was not optimum, but see Appendix B.
- Bearing in mind the discussion in Appendix B, the limited experiments conducted here have shown that basing the LDA transform on full frame rate data has the potential for better performance compared to transforms based on data which has previously undergone VFR analysis.

Overall, LDA has produced some encouraging results from a very limited study.

References

[1] J S Bridle, M D Brown and R M Chamberlain, "A One-Pass Algorithm for Connected Word Recognition", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris, 1982, pp899–902.

[2] L Gillick and S J Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, 23–26 May, pp532–535, 1989.

[3] J N Holmes, "The JSRU Channel Vocoder", IEE Proceedings, vol 127, Part F, number 1, February 1980, pp 53–60.

[4] M J Hunt, "An Introduction to Linear Discriminant Analysis", JSRU Research Report no 1007, 1978.

[5] M J Hunt and C Lefébvre, "Distance Measures for Speech Recognition", National Aeronautical Establishment, Canada, NAE–AN–57, NRR No. 30144, March 1989.

[6] K-F Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System", PhD Thesis, Carnegie Mellon University, 1988.

[7] D S Pallett, W M Fisher, and J G Fiscus "Tools for the Analysis of Benchmark Speech Recognition Tests", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Albuquerque, New Mexico, pp97–100, 1990.

[8] D B Paul, "The Lincoln Robust Continuous Speech Recogniser", ICASSP 89, Glasgow, Scotland, pp449–452, 1989.

[9] S M Peeling and K M Ponting, "Speaker Dependent Speech Recognition Experiments Using Alternative Front Ends With Variable Frame Rate Analysis", RSRE Memo 4389, 1990.

[10] S M Peeling and K M Ponting, "Preliminary Results On The Use of Linear Discriminant Analysis in the *ARM* Continuous Speech Recognition System", RSRE Memo 4511, 1991.

[11] L C W Pols, "Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words", Academische Pers B.V., Amsterdam, 1977.

[12] K M Ponting and S M Peeling. "Word Transition Penalties in the *ARM* Continuous Speech Recognition System", RSRE Memo 4362, 1990.

[13] K M Ponting and M J Russell, "The ARM Project: Automatic Recognition of Spoken Airborne Reconnaissance Reports". Proceedings of 'Military and Government Speech Tech 89', Arlington VA, 13- 15 November, 1989, pp223–227.

[14] M J Russell, "The Speaker Independent ARM Continuous Speech Recognition System", RSRE Memo 4473, 1992.

[15] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle, R K Moore,I Galiano and P Howell, "The ARM Continuous Speech Recognition System" Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Albuquerque, New Mexico, 1990, pp69–72.

Appendix A The Use Of Word Transition Penalties

Several experiments were conducted to investigate the effect of word transition penalties on recognition performance. These were not intended as an exhaustive study into the effect of word penalties; rather they were intended to gauge what, if any, performance improvement was possible. To this end four experiments were conducted which covered both data sets and the different transform matrices:-

- data which had previously undergone VFR analysis.
  - single frame transform
  - three frame transform
- full rate data with VFR analysis applied to the transformed data
  - three frame transform
  - three frame transform with forced differences

The word and phoneme errors for the different types of transform matrices are shown in Figures 7 and 8.

From this it is clear that a significant ($p << 0.0001$) improvement in word recognition performance can be obtained (in each case) by using word penalties of between 10 and 30. All later recognition experiments used word transition penalties of 20, which gave the best performance in these initial experiments.

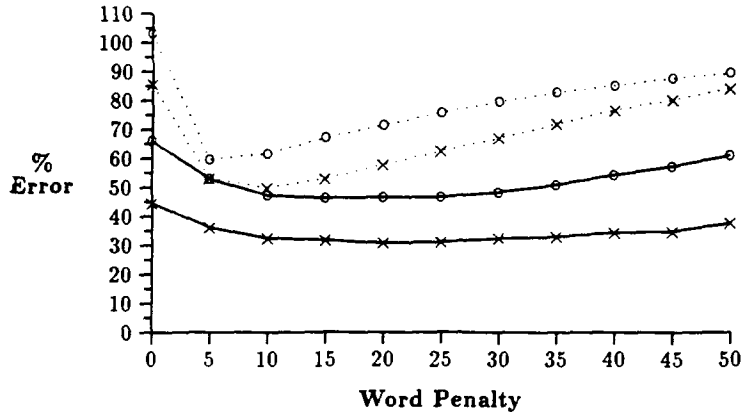

Figure 7: Word (solid line) and phoneme (dotted line) errors for various word transition penalties on speaker independent data with VFR threshold of 900 and LDA transform applied. Single frame transform with 17 elements retained (o) and three frame transform with 19 elements retained (×).

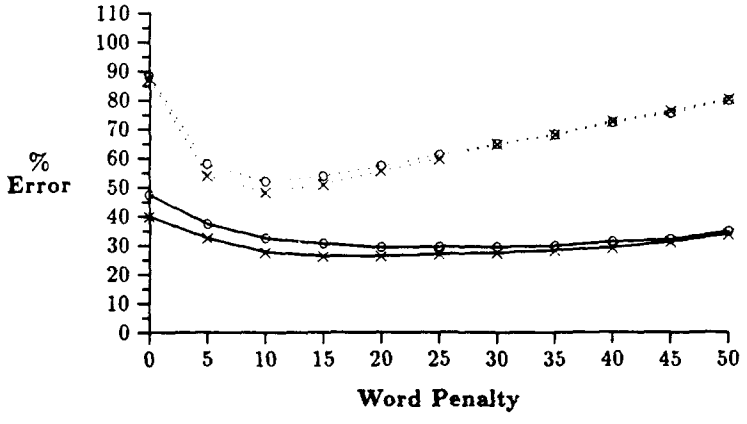

Figure 8: Word (solid line) and phoneme (dotted line) errors for various word transition penalties on full rate speaker independent data with three frame LDA transform and 18 elements retained and VFR threshold of 5. The transform matrices were created without forced differences (o) and with forced differences (×).

Appendix B Comparison Of LDA And MFCC Results

It is difficult to do a direct comparison between the MFCC results quoted by Russell in [14] and the LDA results reported here. There are a number of reasons for this:-

- the feature vectors used for reestimation were of differing sizes.
- the varying degree of data reduction obtained by using VFR analysis.
- differences were used in both cases but they were not always taken over the same time interval.
- VFR analysis was applied at different stages in the pre-processing cycle.

The above list is not exhaustive but serves to show the range of difficulties encountered when trying to do a comparison.

The results reported in [14] are summarised in Table 7 for ease of comparison. Brief descriptions of the different versions of the SI-ARM system are given below; for a full description see [14]. They all use the speaker independent data in the *ARM* system with triphone models together with training and test sets as described in this memo. Each new version of the SI-ARM system builds on the previous one and hence only features which change are described.

- SI-ARM version 1
  This is the "baseline" speaker independent *ARM* system which uses 16 MFCCs with VFR analysis applied (with a threshold of 350) after the cosine transformation. The feature vector contains 18 elements made up of the first 16 cosine coefficients, the average SRUbank channel amplitude and the VFR count.
- SI-ARM version 2
  This differs from version 1 in the ordering of the applications of VFR analysis and cosine transformation. Here, the cosine transformation was applied after the VFR analysis (with a threshold of 1100). Note that this threshold was retained in the later experiments.
- SI-ARM version 3
  The number of cosine coefficients was reduced to 8 but the mean channel amplitude and VFR count were both retained.
- SI-ARM version 4
  Time difference, or "delta cepstrum", information was included. That is, the difference between the feature vectors at $t \pm 1$ was appended to give an output vector containing 20 elements.

- SI-ARM version 5
  A word transition penalty of 30 was used with SI-ARM version 4.

- SI-ARM version 5*
  This version is not explicitly defined in [14] but it is fully described there. It is similar to SI-ARM version 5 but the VFR analysis is applied (with a threshold of 600) after the cosine transformation (which includes delta cepstrum information). This results in an output vector containing 19 elements. The same word transition penalty is used (namely 30).

SI-ARM Version No.	Vector Size	% Frames Remaining	% Word Wrong	% Word Errors
1	18	43.5	28.3	58.1
2	18	35.6	29.8	61.4
3	10	35.6	26.1	51.0
4	20	35.6	17.4	36.1
5	20	35.6	18.2	27.5
5*	19	50.1	17.0	27.2

Table 7: Recognition results obtained using all versions of the SI-ARM system on the test set.

The results from SI-ARM versions 5 and 5* are slightly worse than the best reported here. It should be emphasised that a lot of experimentation, making use of expert knowledge, was performed in order to achieve those results – in contrast to the ones reported here where a fairly simple LDA system was employed.

For all the reasons given above a direct comparison between the two sets of results is difficult. Some of the main points are listed below:–

- It was shown in [14] that similar results are obtained whether the VFR analysis is applied before or after the cosine transformation (cf results for SI-ARM versions 5 and 5*). The (admittedly somewhat limited) results in this memo have shown that it is preferable to apply the LDA transform prior to the application of VFR analysis.

- From a consideration of output vector size then LDA is "better". Performance on a par with SI-ARM versions 5 and 5* has been obtained using as few as 13 features in the output vector. No experiments were conducted into the possibility of using even fewer.

- The early LDA experiments were based on data created using a VFR threshold of 900 rather than the 1100 used in the SI-ARM experiments. The two different

thresholds did not produce significantly different results in [14] so it is difficult to explain the improvement obtained here by using "VFR on VFR".

- The best results reported in [14] were obtained from using VFR thresholds which resulted in about 35% or 50% of the original frames being retained (depending on the order of VFR and MFCC calculation). For LDA, optimum performance is around the 50% point irrespective of the number of features retained.

- If one just considers the frame rate reduction then SIA-ARM version 5 is "better" than LDA . However since 20 features are present in the output vector it is not clear whether this is more computationally efficient than LDA with a 46% reduction and 18 features (which gave the best performance reported here). Similar performance was also obtained here using 13 features with 50% reduction.

It is still valid to claim that LDA has produced some encouraging results from a very limited study.

INTENTIONALLY BLANK

REPORT DOCUMENTATION PAGE

DRIC Reference Number (if known) .....

Overall security classification of sheet .....UNCLASSIFIED.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the field concerned must be marked to indicate the classification eg (R), (C) or (S).

Originators Reference/Report No.	Month	Year
MEMO 4512	JANUARY	1992

Originators Name and Location
RSRE, St Andrews Road
Malvern, Worcs WR14 3PS

Monitoring Agency Name and Location

Title
THE USE OF LINEAR DISCRIMINANT ANALYSIS IN THE *ARM* CONTINUOUS SPEECH RECOGNITION SYSTEM

Report Security Classification	Title Classification (U, R, C or S)
UNCLASSIFIED	U

Foreign Language Title (in the case of translations)

Conference Details

Agency Reference	Contract Number and Period

Project Number	Other References

Authors	Pagination and Ref
PEELING, S M; PONTING, K M	25

Abstract

Linear discriminant analysis is used to generate speech data transformations. This transformed data is then used within the *ARM* continuous speech recognition system. Experiments are described using transformed data in conjunction with variable frame rate analysis and word transition penalties. Speaker independent results are reported which are as good as the best obtained previously using cosine transformations and variable frame rate analysis. The two sets of results are compared and commented on.

Abstract Classification (U,R,C or S)
U

Descriptors

Distribution Statement (Enter any limitations on the distribution of the document)
UNLIMITED

INTENTIONALLY BLANK